Gibbs sampler for background discrimination in particle physics

Federico Colecchia*
Department of Physics and Astronomy, University College London

Background properties in experimental particle physics are typically estimated from large collections of events. This usually provides precise knowledge of average background distributions, but inevitably hides fluctuations. To overcome this limitation, an approach based on statistical mixture model decomposition is presented. Events are treated as heterogeneous populations comprising particles originating from different processes, and individual particles are mapped to a process of interest on a probabilistic basis. When used to discriminate against background, the proposed technique based on the Gibbs sampler allows features of the background distributions to be estimated directly from the data without training on high-statistics control samples. A feasibility study on Monte Carlo is presented, together with a comparison with existing techniques. Finally, the prospects for the development of the Gibbs sampler into a tool for intensive offline analysis of interesting events at the Large Hadron Collider are discussed.

Keywords: 29.85.Fj; High Energy Physics; Particle Physics; LHC; background discrimination; Gibbs sampler; Markov Chain Monte Carlo.

1 Introduction

Background discrimination in particle physics is usually performed by identifing events that are more likely to contain a physics process of interest, the primary goal being rejection of contributions from uninteresting processes that mimic the signal and thus make its extraction and measurement more complicated. Traditional approaches achieve this goal by focussing on entire events, comparing kinematic and topological properties with reference distributions usually obtained from control samples.

This article presents a novel approach that builds on a population-based view of particle physics events, which are treated as collections of particles originating from different physics processes such as a hard scattering of interest as opposed to background. The main goal is not to identify events that are more likely to contain a given signal process, but rather to process collections of particles and map each of them to different processes on a probabilistic basis.

This is achieved by applying mixture model decomposition techniques [1], which are well established in statistics and which have been used in other disciplines to solve formally-similar problems. In this formulation, events are treated as heterogeneous statistical populations comprising particles whose kinematics and topology reflect the process they originated from: an input collection of particles is processed using an iterative algorithm

*Department of Physics and Astronomy, University College London, Gower Street, London, WC1E 6BT. Tel. +44 20 7679 3425. Email: federico@hep.ucl.ac.uk

that assigns individual particles a probability for them to be associated with a hard scattering of interest as opposed to background.

This contribution describes a feasibility study for the use of the Gibbs sampler for background discrimination at the Large Hadron Collider (LHC).

The Gibbs sampler was originally proposed with regards to Bayesian restoration of images [2] and has subsequently been used to solve different problems in a number of disciplines. The Gibbs sampler and, more generally, Markov Chain Monte Carlo (MCMC) techniques have also been recently applied to the study of the Cosmic Microwave Background radiation [3]. As far as particle physics is concerned, the last few years have witnessed a renewed interest in Bayesian numerical methods in the context of a general effort to take advantage of the flexibility of Bayesian statistics for data analysis [4] [5] [6], in addition to the use of MCMC methods with reference to specific optimization problems [7].

In the context of this study, the Gibbs sampler iterates on an input collection of particles in order to decompose the corresponding mixture into particles originating from a hard scattering of interest as opposed to background, and individual particles are mapped to signal or background on a probabilistic basis. Results obtained using the Gibbs sampler on a collection of ~ 600 simulated particles originating from a hard scattering and from background are reported in this article, together with results from additional toy Monte Carlo studies.

Although the proposed approach can be seen as a solution to a classification problem where individual particles are considered as opposed to entire events, the technique presented here differs significantly from the simple application of existing classifiers to individual particles. Traditional classification techniques in particle physics in fact rely on supervised algorithms, which are first trained on high-statistics control samples, on which they "learn" distinctive features of signal and background signatures, and which are subsequently applied to a test data set. This assumes that background properties estimated from a high-statistics training sample also reliably describe background activity in the data set under study, which often comprises a significantly-lower number of particles than the control sample that was used to train the classifier.

However, the above assumption is often wrong: different events in particle physics can in fact look very different from one another even when the underlying physics processes are the same, and statistical fluctuations can be non-negligible in low-statistics data sets. When classification is performed using traditional supervised algorithms, fluctuations are usually not taken into account, since training typically relies on high-statistics control samples. The Gibbs sampler as presented in this article can instead also be operated in unsupervised mode: this allows the algorithm to estimate signal and background probability density functions (PDFs) directly from the data, and to map individual particles to signal or background based on those estimates instead of predefined high-statistics templates.

From a broader perspective, this contribution describes a set of techniques that exploit the flexibility of the Gibbs sampler to estimate signal and background properties directly from the data, and to assign individual particles a probability for them to originate from either process. This article does not present an established method for background discrimination at the LHC, but rather discusses results of a feasibility study for the use of the Gibbs sampler in different configurations, the primary objective being the decomposition of an input collection of particles into signal and background-associated subpopulations. While the algorithm is not here proposed as a tool for intensive offline analysis of individual interesting events at the LHC, the total number of particles chosen for this proof of concept is in line with typical charged particle multiplicities corresponding to LHC operating conditions as of July 2011. For this reason, the present implementation of the algorithm is a

promising starting point with a view to developing techniques based on the Gibbs sampler for the analysis of individual LHC events at the particle level.

2 The algorithm

This novel algorithm for background discrimination is presented with reference to the general problem of decomposing a collection of particles from high-energy particle collisions into subpopulations associated with different underlying physics processes and described in terms of different probability distributions.

The input data set thus consists of a mixture of particles, some of which originated from a hard scattering of interest, others from a background process. Provided that the corresponding subpopulations can be characterized sufficiently well in terms of their kinematic or topological properties, it is in principle possible to ask, for each particle, what the probability is for it to originate from signal as opposed to background. In particular, the Gibbs sampler can estimate such probabilities by iteratively sampling from the subpopulation PDFs.

Although mixture model decomposition techniques based on the Gibbs sampler are well-established in statistics, classical algorithms are not suited for a direct application to the problem at hand without modification. Classical mixture models in fact usually rely on a parametric formulation that requires the shapes of the PDFs of the relevant subpopulations to be known a priori. However, our previous studies suggest that this assumption may be responsible for non-negligible bias when the Gibbs sampler is applied to the problem at hand. The latter bias ultimately relates to the assumption that PDF functional forms obtained from high-statistics control samples are also appropriate to describe the corresponding probability distributions in the input data set, which often comprises a significantly-lower number of particles, and in which statistical fluctuations may not be negligible.

For this reason, the statistical model for the Gibbs sampler is here formulated in more general terms with respect to classical mixture models used in statistics, without requiring knowledge of the PDF functional forms of the relevant subpopulations. The model is based on the following PDF:

$$\sum_{j=1}^{k} \alpha_j f_j(x) \tag{1}$$

with subpopulation fractions α_j ("mixture weights") satisfying:

$$\sum_{j=1}^{k} \alpha_j = 1,$$

The variable x can correspond to particle pseudorapidity η, a kinematic variable related to the polar angle of the particle, or p_T i.e. the transverse momentum of the particle with respect to the beam direction.

The subpopulation PDFs f_j are defined in terms of regularized histograms of x, as described in section 3. The symbol φ_j is used to denote the estimate of the generic subpopulation PDF f_j throughout the text.

Given the above mixture of probability distributions and given a set of observations $\{x_i\}_{i=1,...,N}$, the problem of clustering the observations into k groups by probabilistically associating each of them with a distribution of origin has been solved numerically in a Bayesian framework using MCMC techniques, in particular using the Gibbs sampler [1].

The pseudocode of the algorithm is reported below. The value of variable v at iteration t is indicated with $v^{(t)}$ throughout.

1. **Initialization:** Choose $\underline{\alpha}^{(0)} = \{\alpha_j^{(0)}\}_j$ and $f_j^{(0)}{=}\varphi_j^{(0)}$, $j = 1, ..., k$ as described in section 3.

2. **Iteration t:**

   (a) Generate the allocation variables $z_{ij}^{(t)}$, $i = 1, ..., N$, $j = 1, ..., k$ based on prob-

abilities $P(z_{ij}^{(t)} = 1|\alpha_j^{(t-1)}, \varphi_j^{(t-1)}, x_i)$ proportional to $\alpha_j^{(t-1)} f(x_i|\varphi_j^{(t-1)})$. The quantity $z_{ij}^{(t)}$ equals 1 when observation i is mapped to distribution j at iteration t, and 0 otherwise. In general, the variables $z_{ij}^{(t)}$ depend both on the mixture weights α_j and on the estimates φ_j of the subpopulation PDFs from the previous iteration.

(b) Generate $\underline{\alpha}^{(t)}$ from the probability density function of $\underline{\alpha}$ given $\underline{z}^{(t-1)} = \{z_{ij}^{(t-1)}\}_{ij}$, $\rho(\underline{\alpha}|\underline{z}^{(t-1)})$. Knowledge of which particles are mapped to process j at iteration $t-1$ makes it possible to generate the subpopulation fractions $\underline{\alpha}$ at iteration t.

(c) Obtain an updated estimate of the subpopulation PDFs from the data $\underline{x}$ based on knowledge of which particles are mapped to subpopulation j at iteration $t-1$. Details are provided in section 3.

A specific choice for the function ρ and a way to obtain updated estimates of the subpopulation PDFs f_j are described in section 3 with reference to a Monte Carlo study.

The central idea of the algorithm is the following: the better the observations $\{x_i\}_i$ are mapped to the subpopulations $j = 1, ..., k$, the more accurate the estimates of $\rho(\underline{\alpha}|\underline{z})$ and of the subpopulation PDFs φ_j. Once some correct values of z_{ij} are found, $\rho(\underline{\alpha}|\underline{z})$ and φ_j begin to roughly reflect the correct distributions, which in turn leads to additional correct mappings z_{ij} to be found in subsequent iterations.

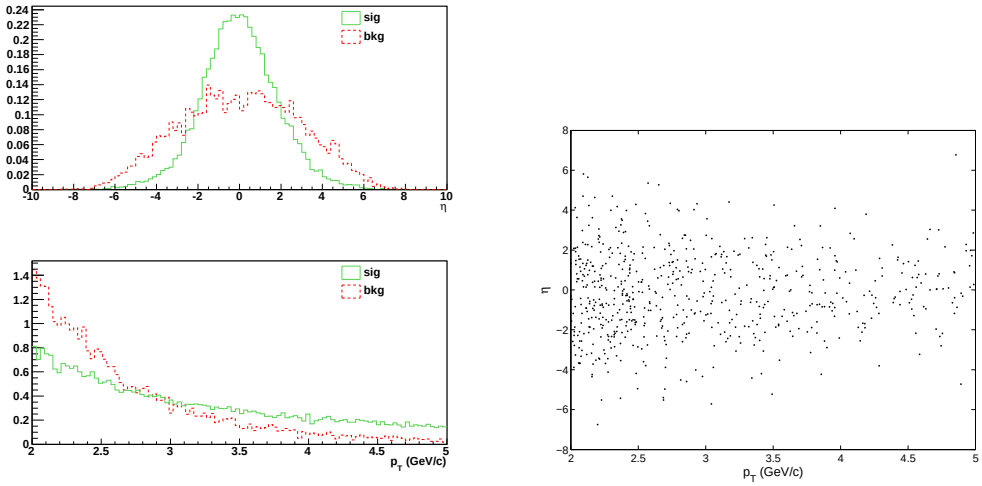
The Gibbs sampler can be operated both in supervised and unsupervised mode. The above pseudocode corresponds to the "unsupervised Gibbs sampler", where updated estimates of subpopulation PDFs are obtained at each iteration, as indicated at step (c). On the other hand, when step (c) is removed from the pseudocode, particles are mapped to signal and background based on the subpopulation PDFs provided at initialization, and the algorihtm is operated in supervised mode ("supervised Gibbs sampler"). The primary objective of this contribution is to propose the use of the Gibbs sampler in different configurations in order to:

- Obtain estimates φ_j of the subpopulation PDFs from the input data set.
- Estimate the subpopulation weights α_j. In the context of this study, this corresponds to estimating the fractions of background and signal particles contained in the input data set.
- Assign individual particles a probability for them to originate from a given process based on the subpopulation PDFs estimated at step (a) as opposed to relying on high-statistics templates. In the context of this study, this allows classification of individual particles into signal and background.

This article reports the results of a feasibility study for one possible combination of the supervised and unsupervised Gibbs sampler that was chosen with a view to minimizing systematic uncertainties. However, other modes of operation of the algorithm may be appropriate for specific applications, and the exact configuration of the Gibbs sampler may have to be tailored to the analysis at hand.

3 Monte Carlo study

The Gibbs sampler was applied to a Monte Carlo data set generated using Pythia 8.140 [8] [9], obtained superimposing $gg \to t\bar{t}$ signal events from pp interactions at $\sqrt{s} = 14$ TeV to soft QCD interactions, so called Minimum Bias events, in order to simulate background. The signal process was chosen in order to illustrate the use of the algorithm for background discrimination at the particle level. Further studies will be needed in order to extend these results beyond the proof of concept



(a) (b)

Figure 1: Generator-level η (top left) and p_T (bottom left) distributions for signal (solid green histograms) and background particles (dashed red histograms) with 2 GeV/c $< p_T <$ 5 GeV/c from the high-statistics control sample. The distributions correspond to a total number of $\sim 33,000$ particles and are normalized to unit area. The corresponding two-dimensional distribution is displayed on the right-hand side of the figure.

presented in this article, and to assess the potential of the Gibbs sampler for background discrimination in the context of specific analyses at the LHC.

The Gibbs sampler was run over a collection of charged particles with 2 GeV/c $< p_T <$ 5 GeV/c, and individual particles were assigned a probability for them to originate from signal as opposed to background based on their η and p_T values.

The pseudocode of the algorithm used for this application is shown below. Subscripts *sig* and *bkg* relate to signal and background, respectively.

1. **Initialization:** Set $\alpha_{bkg} = \alpha_{bkg}^{(0)} = 0.5$, $f_j = \varphi_j^{(0)}$. Initial conditions for the estimates $\varphi_j^{(0)}$ of the subpopulation PDFs $f_j^{(0)}$ are given by regularized η and p_T distributions from the high-statistics control sample, as described in section 3.1.

2. **Iteration t:**

   (a) Generate $z_{ij}^{(t)}$ for all particles ($i = 1, ..., N$) and distributions ($j = 1, 2$ corresponding to background and signal, respectively) according to $P(z_{ij}^{(t)} = 1|\alpha_j^{(t-1)}, \varphi_j^{(t-1)}, x_i) \propto \alpha_j^{(t-1)} f_j(x_i|\varphi_j^{(t-1)})$, where $\alpha_1 = \alpha_{bkg}$, $\alpha_2 = 1 - \alpha_{bkg}$.

   (b) Set $\alpha_j^{(t)} = \sum_i z_{ij}^{(t-1)}/N$, $\forall j$. This corresponds to the simplest choice of setting $\rho(\alpha_j|\underline{z}^{(t-1)}) = \delta(\alpha_j - \sum_i z_{ij}^{(t-1)}/N)$ for the probability density function of $\underline{\alpha}$ given $\underline{z}$.

   (c) Obtain updated estimates of the subpopulation PDFs by regularizing the η and p_T distributions corresponding to particles mapped to the relevant subpopulation at iteration $t-1$, i.e. based on $z_{ij}^{(t-1)}$.

In general, the functions f_j are the joint PDFs for η and p_T corresponding to background ($j = 1$)

and signal ($j = 2$) particles. This study is restricted to charged particles with 2 GeV/c $< p_T <$ 5 GeV/c, which makes it possible to neglect the correlation between η and p_T as a first approximation. For this reason, the joint PDFs take the form $f_{sig/bkg} = f^{(\eta)}_{sig/bkg} f^{(p_T)}_{sig/bkg}$, and obtaining updated estimates of the subpopulation PDFs reduces to regularization of one-dimensional histograms, as described in the following.

As for the number of iterations to be used with the Gibbs sampler, no rule is documented in the literature, and the choice is generally problem-dependent. The number of iterations was set to 1,000 in this study, and probabilities were averaged over the last 100 iterations. Multiple runs were performed under different conditions and using different numbers of iterations to make sure the algorithm converged. The issue of whether the stationary distribution has been reached by the sampler has no clear solution in the literature, and established practice relies on visual inspection of trace plots [1].

In order to obtain initial conditions $\varphi^{(0)}_j$ for the subpopulation PDFs, a Monte Carlo data set was used containing a total of about 33,000 charged particles in the kinematic range 2 GeV/c $< p_T <$ 5 GeV/c. In addition to estimation of $\varphi^{(0)}_j$, this high-statistics control sample was also used to guide the histogram regularization procedure as described in the following. Figure 1 (a) shows the η and p_T distributions for signal (solid green histograms) and background particles (dashed red histograms).

As anticipated, one of the goals of the Gibbs sampler is to estimate the background PDFs directly from the input collection of particles. In other words, the algorithm can classify particles into signal and background without using any predefined background templates: the background PDFs that are estimated by the algorithm thus reflect the specific background conditions in the input data set, which can be different from "average" conditions estimated from a high-statistics control sample.

The Gibbs sampler basically tries to uncover a signal and a background subpopulation in the input collection of particles based on the data and on initial conditions for the subpopulation PDFs. The results presented in this article relate to an input data set comprising 636 charged particles in the above-mentioned kinematic region 2 GeV/c $< p_T <$ 5 GeV/c, out of which 481 originate from a signal hard process and 155 from background, corresponding to a fraction of background particles of $\sim$ 24%. The total number of particles in the input data set is in line with typical charged particle multiplicities at the LHC as of July 2011.

The signal and background η and p_T distributions corresponding to the Monte Carlo data set used in this study are shown in figure 2. The solid green (dashed red) histograms in the upper panel display the signal (background) η distributions, normalized to unit area. The corresponding p_T distributions are given in the lower panel.

Figure 3 shows an example of the $x = \eta$ distribution for particles mapped to the signal subpopulation at a given step of the Gibbs sampler. As opposed to assuming a functional form for the PDF and fitting a function to the histogram, the histogram is regularized i.e. the subpopulation PDF is obtained by means of spline interpolation of the histogram contents, as further discussed in the following. The superimposed curve on the figure corresponds to the regularized histogram, and is used by the Gibbs sampler as an estimate of the corresponding subpopulation PDF f_j.

This approach gives the algorithm more flexibility in terms of estimating the subpopulation PDFs directly from the input data set, and results in a significant reduction of the biases observed in our previous studies.

[1] In general, convergence of a MCMC time series can be linked to the number of draws from the sequence and to the positive auto-correlation between adjacent members [10] [11]. In particular, the standard deviation of the sequence with respect to the stationaty value scales as the inverse square root of the number of draws, with a multiplicative factor accounting for auto-correlation.

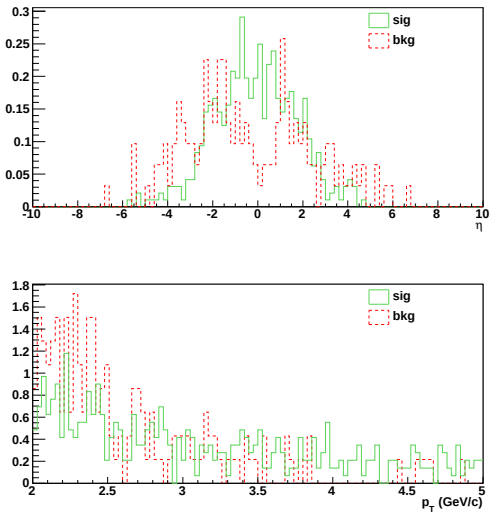


Figure 2: Particle η and p_T distributions from the Monte Carlo input data set used in this study. Solid green and dashed red histograms correspond to signal and background, respectively. Distributions are normalized to unit area.

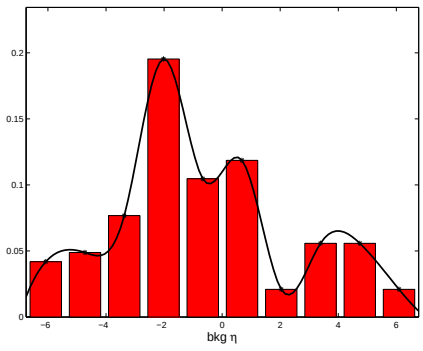


Figure 3: Example of the pseudorapidity η distribution of particles mapped to the background subpopulation at a given step of the Gibbs sampler. The superimposed curve is the result of the regularization procedure described in the text, and is used by the Gibbs sampler as an estimate of the corresponding subpopulation PDF.

3.1 Regularization

Step (c) in the pseudocode shown in section 3 requires iterative PDF updates based on the current mapping of individual particles to the different subpopulations. This operation is performed when the Gibbs sampler is operated in unsupervised mode, and generally has a significant impact on the results, as described below.

In general, a histogram representing the η or p_T distribution of particles mapped to signal or background at a given iteration cannot be used directly as an estimate of the underlying PDF due to the presence of statistical fluctuations. In fact, a simple spline interpolation of the histogram contents, e.g. by means of cubic interpolation, generally leads to fluctuations being mistaken for genuine features of the underying PDFs, and the Gibbs sampler is typically unable to converge. This is similar to what happens when observed distributions in particle physics are unfolded in order to "remove" detector effects: regularization techniques have been investigated in detail in that context, for instance as a way to use a priori information about the underlying distributions in order to get rid of spurious oscillatory components (see e.g. [12]).

The complexity of the histogram regularization procedure that was used in this study was intentionally kept minimal in order to avoid the introduction of additional complications that may obscure the response of the algorithm at this stage of its development. The possible incorporation of more developed regularization methods is left for future studies.

A priori information about the signal and background PDFs was obtained from the high-statistics control sample. When subpopulation PDFs are estimated during the execution of the Gibbs sampler in unsupervised mode, a "regularization window" is applied to the η histograms in order to get rid of outliers: in other words, a spline interpolation of the histogram contents is obtained using only the part of the histogram that lies between a minimum and a maximum η value, which leads to extreme statistical fluctuations on the tails of the distribution to be excluded. It is worth noticing that assuming a functional form for the PDF and fitting it to the histogram would effectively produce a similar result, i.e. would reduce the impact of outliers on the estimated PDF. However, as anticipated, that approach was observed to introduce significant bias in previous studies, and was thus abandoned in favour of the non-parametric statistical model presented in (1), where subpopulation PDFs are defined as the output of a histogram regularization procedure without reference to any predefined functional form. Figure 4 shows signal and background η distributions from the high-statistics control sample, with superimposed arrows indicating the "regularization window". The maximum $|\eta|$ value was set to $|\eta| = 5$ ($|\eta| = 7$) for signal (background) in this study.

On top of this, again with a view to getting rid of extreme statistical fluctuations when regularizing histograms, boundary conditions were introduced on the η and p_T PDFs, constraining the value of f_j to points chosen based on control sample distributions: in particular, the signal (background) η PDF was constrained to 0 when $|\eta| > 5$ ($|\eta| > 7$), and signal (background) p_T PDFs were constrained at 2 GeV/c and 5 GeV/c to 0.7 (1.2) and 0.1 (0) (see figure 1(a)).

Results were found to be stable with respect to reasonable changes to the regularization contraints reported above.

3.2 Gibbs sampler configuration

As anticipated, the Gibbs sampler can be operated both in supervised and unsupervised mode, and this article presents results obtained using a particular configuration of the algorithm, corresponding to a combination of the two modes of operation chosen to minimize systematic uncertainties.

As already pointed out in section 2, the Gibbs sampler may be used to process an input collection of particles in order to obtain one or more of the following results:

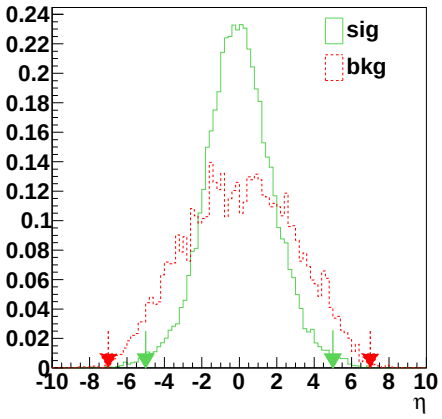


Figure 4: Illustration of the "regularization window" used in this study for histogram regularization. The histograms correspond to signal (solid green) and background (dashed red) η distributions from the high-statistics control sample. The position of the solid green and dashed red arrows correspond to the "regularization window": at each iteration of the Gibbs sampler, only the part of the distribution that lies between the arrows is used for spline interpolation, which makes the results robust against outliers. Additional details are given in the text.

1. Estimate the subpopulation PDFs from the input data set.

2. Estimate the fraction of particles associated with a given process present in the input data set, e.g. the fraction of background particles.

3. Assign individual particles a probability for them to originate from a given process, such as a hard scattering of interest as opposed to background, based on subpopulation PDFs estimated at step (1) as opposed to relying on predefined templates.

Depending on the objective, it may be appropriate to run the algorithm in different modes.

For instance, the histogram regularization procedure that is used to obtain iterative estimates of the subpopulation PDFs with the unsupervised Gibbs sampler directly leads to a bias on the mixture weights. For this reason, the unsupervised Gibbs sampler may not be the most appropriate tool to estimate the fraction of particles in the input data set originating from a given process, such as the fraction of background particles. On the other hand, when the algorithm is operated in unsupervised mode, it can obtain data-driven estimates of the subpopulation PDFs, a result that cannot be achieved using traditional supervised techniques.

One possible combination of supervised and unsupervised Gibbs sampler is here proposed where:

1. The supervised Gibbs sampler is first used to estimate the mixture weights. In the two-subpopulation scenario described in this study, goal (2) above corresponds to estimating the fraction of background particles contained in the input data set. The initial conditions correspond to a fraction of background particles of 0.5, and the subpopulation PDFs at this stage are fixed to the estimates provided by the high-statistics control sample. The corresponding results are presented and discussed in section 3.2.1.

2. The Gibbs sampler is run again on the input data set. This second run is performed in

unsupervised mode, i.e. subpopulation PDFs are now updated at each iteration, starting from initial conditions corresponding to regularized distributions from the high-statistics control sample. On the other hand, the mixture weights are kept fixed to the results from the previous step.

As for assigning individual particles in the input data set a probability for them to originate from signal as opposed to background, the most appropriate approach may again depend on the specific application. In general, probabilities may be assigned directly using the unsupervised Gibbs sampler at step (2) above, as done in this study, or an additional run of the supervised Gibbs sampler may altenatively be added after the previous two, with fixed PDFs given by the estimates from step (2). Further studies will be necessary in order to better understand the classification performance of the Gibbs sampler under different conditions and to guide this choice.

Results obtained running the supervised and unsupervised Gibbs sampler as described above on the Monte Carlo data set used in this study are reported and discussed in the following sections.

3.2.1 Supervised Gibbs sampler

The supervised Gibbs sampler was primarily used in this study to estimate the mixture weights of the statistical model, i.e. the fraction of background particles contained in the input data set. Figure 5 shows the corresponding estimates over the last 100 iterations. The solid green and dashed red curves correspond to the estimated fractions of signal and background particles, respectively. The solid green (dashed red) horizontal line indicates the signal (background) true value from the simulation, while the dash-dot line corresponds to the initial conditions for the mixture weights.

These results indicate a bias on the estimated mixture weights. On the other hand, using the algorithm in supervised mode to estimate the subpopulation mixture weights would introduce a dependence on the regularization parameters due to the histogram regularization procedure discussed in section 3.1, which is not desirable. This study has pointed to the following possible reasons for the observed bias on the mixture weights when the Gibbs sampler is operated in supervised mode:

- Predefined PDF templates: when the algorithm is run in supervised mode, the subpopulation PDFs are not estimated from the data at each iteration, and particles are mapped to signal or background based on the PDF templates obtained from the high-statistics control sample. This can lead to a bias on the estimated mixture weights, which may be different for different input data sets. This is an irreducible source of bias on the supopulation mixture weights when the Gibbs sampler is operated in supervised mode.

- “Sampling bias": Even when the PDF templates obtained from the high-statistics control sample provide a good description of the signal and background PDFs in the input data set, the use of those templates on a lower-statistics data set can still lead to a bias on the mixture weights. In general, the RMS of a histogram obtained sampling from a probability distribution in fact increases with the number of samples, and it may thus be necessary to correct the “high-statistics" PDF templates before applying them to a lower-statistics input data set in order to avoid bias. Future studies will investigate possible ways to mitigate this effect.

In order to make sure that the bias in figure 5 is related to information provided to the Gibbs sampler based on the high-statistics control sample rather than to an intrinsic bias in the algorithm, additional runs on toy Monte Carlo samples were performed, as described in appendix A. In particular, figure 6 displays the estimated mixture weights obtained running the supervised Gibbs sampler on a toy Monte Carlo data set with subpopulation

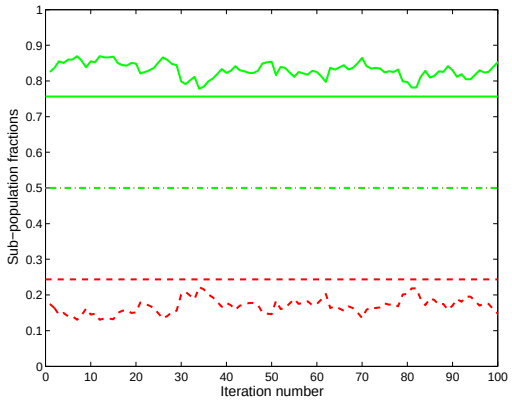


Figure 5: Mixture weights obtained running the supervised Gibbs sampler on the input Monte Carlo data set. Results from the last 100 iterations are shown. The solid green (dashed red) curve denotes the estimated fraction of signal (background) particles. The solid green (dashed red) horizontal line indicates the true value for signal (background) from the simulation, and the dash-dot line corresponds to the initial conditions for the mixture weights. The observed bias is discussed in the text.

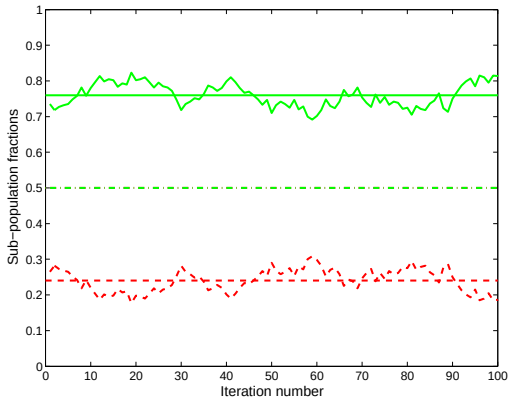


Figure 6: Mixture weights obtained running the supervised Gibbs sampler on a toy Monte Carlo data set, as described in the text. Results from the last 100 iterations are shown. The solid green (dashed red) curve corresponds to the estimated fraction of signal (background) particles. The solid green (dashed red) horizontal line indicates the true value for signal (background) from the toy Monte Carlo, and the dash-dot line corresponds to the initial conditions for the mixture weights.

PDFs kept fixed at truth information. In this case, no significant bias is expected, and the figure confirms a better agreement between estimated and true mixture weights with respect to the corresponding results in figure 5.

3.2.2 Unsupervised Gibbs sampler

The unsupervised Gibbs sampler was used in this study to estimate the signal and background PDFs from the input data set, while keeping the mixture weights fixed at the results obtained from the previous run of the algorithm in supervised mode.

Figure 7 shows the subpopulation PDFs estimated by the Gibbs sampler on the Monte Carlo input data set. Solid curves correspond to the output of the histogram regularization procedure averaged over the last 100 iterations, superimposed to the true distributions (histograms). The η (p_T) distributions are displayed in the top (bottom) plots, figures on the left-hand (right-hand) side correponding to background (signal). All distributions are normalized to unit area. The bottom panel in each figure shows the corresponding ratio between the relevant subpopulation PDF estimated by the algorithm and truth information.

In addition to obtaining data-driven estimates of the subpopulation PDFs in unsupervised mode, one of the goals of the Gibbs sampler in this application is to assign individual particles a probability for them to originate from a given process, such as a hard scattering of interest as opposed to background. In this study, the latter probabilities were obtained from the same unsupervised run of the algorithm that provided the PDF estimates in figure 7. In general, other choices are possible, such as performing an additional supervised run of the Gibbs sampler with subpopulation PDFs kept fixed at the estimates shown in figure 7.

A first comparison of the classification performance of the algorithm in the configuration chosen for this study with the corresponding performance of existing techniques is described in section 3.3. Additional studies will be required for a detailed assessment of the classification performance of the algorithm corresponding to different operation modes.

The probabilities returned by the Gibbs sampler were validated by comparing the true kinematic distributions with the corresponding ones for particles with $P_{sig} > 0.5$, P_{sig} being the probability for a given particle to originate from the signal process, averaged over the last 100 iterations of the algorithm. Results are shown in figure 8, where histogram bars indicate the η distribution for particles with $P_{sig} > 0.5$, while stars correspond to true distributions.

3.3 Classification performance

Operating the Gibbs sampler as presented in this article is equivalent to using it as a binary classifier. Its performance can thus be quantified using the Receiver Operating Characteristic (ROC) curve, which displays true-positive as a function of false-positive probability. The area under the curve is a number between 0 and 1: the higher its value, the better the classifier is able to discriminate between the two categories (signal and background in this case). The ROC curve of a random classifier would be a straight line along the main diagonal on the true-positive vs false-positive plane ("chance diagonal" [13]).

Figure 9 shows a comparison between the ROC curve obtained using the unsupervised Gibbs sampler on the input collection of particles used for this study and the corresponding curves from different supervised multivariate classification methods using TMVA [14] V04-01-00. The curves are displayed using an equivalent representation in terms of background rejection rate as a function of signal efficiency[2] . The dashed lines refer to

[2] Based on our previous terminology, "background rejection rate" is equivalent to $1 - P_{bkg\to sig}$, and "signal efficiency" corresponds to $P_{sig\to sig}$, where $P_{bkg\to sig}$ ($P_{sig\to sig}$) is the probability for a background (signal) particle to be mapped to the signal subpopulation.

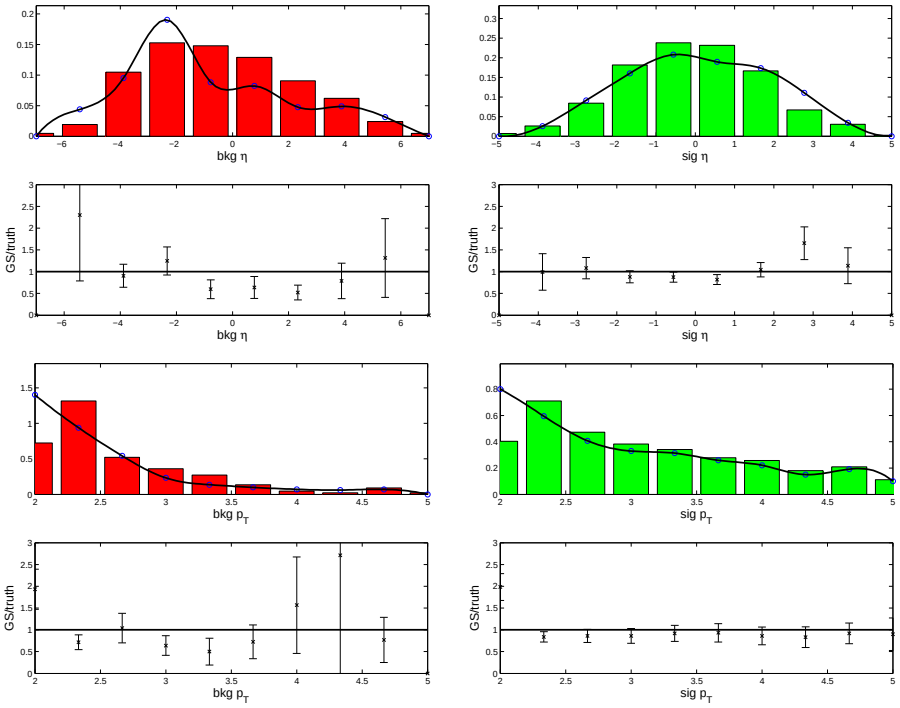


Figure 7: Subpopulation PDFs estimated by the Gibbs sampler in unsupervised mode on the input Monte Carlo data set used in this study, averaged over the last 100 iterations. Top left: background η. Top right: signal η. Bottom left: background p_T. Bottom right: signal p_T. In each subfigure, the upper panel shows truth information (histogram bars) superimposed to the result of the regularization procedure averaged over the last 100 iterations (curve). The lower panels display the ratio between subpopulation PDFs estimated by the algorithm and the corresponding truth-level information.

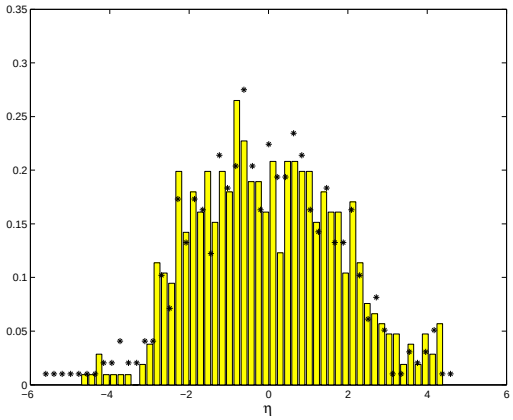


Figure 8: Comparison between η and p_T distributions for particles with $P_{sig} > 0.5$ (histogram bars) and truth (stars). Additional information is given in the text.

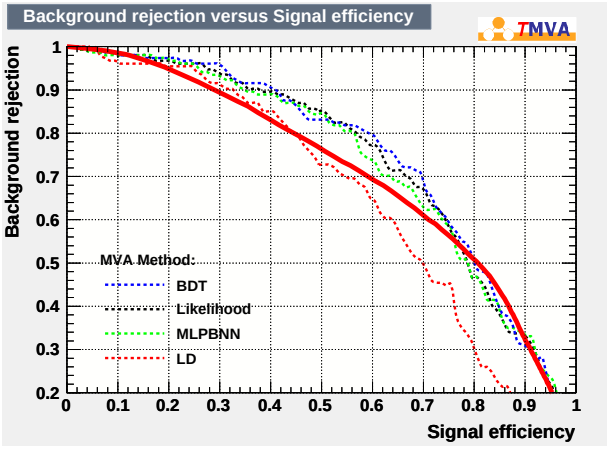


Figure 9: Comparison between the ROC curve obtained using the Gibbs sampler in unsupervised mode, shown here as background rejection rate as a function of signal efficiency, and the corresponding curves obtained using existing supervised classification techniques from TMVA, as described in the text. The solid red line is the curve from the unsupervised Gibbs sampler corresponding to the last 100 iterations. The other curves correspond to TMVA algorithms, namely Boosted Decision Trees (dashed blue), Naive Bayes classification (dashed black), the Neural Network-based classifier MLPBNN (dashed green), and Linear Discriminant (dashed red).

$L(\mathrm{cm}^{-2}\mathrm{s}^{-1})$	$\langle n_{PU} \rangle$	N_{bkg}/N_{sig}
10^{33}	2.3	0.1
10^{34}	23	0.9
10^{35}	230	4.4

Table 1: Average numbers of pile-up particles (second column) expected at different LHC instantaneous luminosities (first column) [15]. A 25 ns bunch crossing is assumed. The third column reports the corresponding ratios between the number of background and signal particles observed in the kinematic region considered in this study.

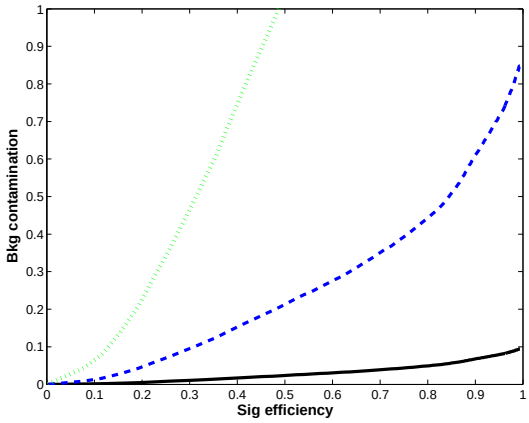


Figure 10: Background contamination as a function of signal efficiency at different LHC instantaneous luminosities (10^{33}cm^{-2}s^{-1} solid black, 10^{34}cm^{-2}s^{-1} dashed blue, 10^{35}cm^{-2}s^{-1} dotted green). Background contamination is defined as the number of misclassified background particles normalized to the number of signal particles. Misclassification probabilities correspond to the ROC curve from the Gibbs sampler in figure 9. A 25 ns bunch crossing is assumed [15].

different TMVA methods[3], namely Boosted Decision Trees (dashed blue), Naive Bayes classification (dashed black), the Neural Network-based classifier MLPBNN (dashed green), and Linear Discriminant (dashed red).

The solid red line corresponds to the Gibbs sampler. The figure suggests that classification performance of the Gibbs sampler is similar to that of existing supervised methods.

It may also be useful to provide a more precise idea of the background rejections and signal efficiencies that can be achieved using the Gibbs sampler corresponding to different LHC instantaneous luminosities[4]. Figure 10 shows estimates of background contamination as a function of signal efficiency at three different LHC instantaneous luminosities. Background contamination is defined as the number of misclassified background particles normalized to the number of signal particles, and is calculated by rescaling the abscissa of the ROC curve by the ratio between the number of background and signal particles in the kinematic region considered in this study, as given in table 1[5]. The three curves correspond to instantaneous luminosities of 10^{33}cm^{-2}s^{-1} (solid black), 10^{34}cm^{-2}s^{-1} (dashed blue), and 10^{35}cm^{-2}s^{-1} (dotted green).

3.4 Additional remarks

The possibility to be operated in unsupervised mode distinguishes the Gibbs sampler from existing multivariate approaches such as those available in ROOT [16] with TMVA. However, the iterative nature of the algorithm leads to a disadvantage with respect to established multivariate algorithms in terms of execution time. The running time of the Gibbs sampler corresponding to 1,000 iterations on the Monte Carlo input data set used in this study was about 20 s on a 2GHz Intel Processor with 1GB RAM.

[3] The algorithms were run using the high-statistics control sample for training and the same collection of particles the Gibbs sampler was run on for testing.

[4] The expected average number of pile-up interactions, i.e. the expected average number of primary vertices in the events is here taken as a measure of background activity for illustrative purposes.

[5] The average numbers of pile-up interactions at different LHC instantaneous luminosities are taken from [15], and correspond to a 25 ns bunch crossing.

However, given the parallelization potential of the Gibbs sampler that was pointed out already in the original paper [2], improvements may be possible in this respect, for example using commodity Graphics Processing Units (GPUs) that are extensively used both in particle physics and in other disciplines for compute-intensive applications.

4 Conclusions and outlook

This contribution has presented a feasibility study for a novel approach to background discrimination in particle physics that builds on a population-based view of events from high-energy particle collisions. Collections of particles are here treated as mixtures of subpopulations associated with different physics processes, and numerical methods well established in statistics for mixture model decomposition are used to assign individual particles a probability for them to originate from a hard scattering of interest as opposed to background. This application of the Gibbs sampler to a classification problem at the particle level allows the development of a flexible suite of tools to estimate background properties directly from the data, e.g. to obtain PDF shapes without relying on templates obtained from high-statistics control samples and without assuming predefined functional forms.

This study has highlighted strength and limitations of the algorithm operated in supervised and unsupervised mode, and a possible combination has been proposed to reduce bias. In general, systematic uncertainties associated with the use of the Gibbs sampler will have to be evaluated in the context of a given analysis.

Understanding in detail how the classification performance of the algorithm in different configurations compares to existing techniques will require further study, as will the possible development of subsequent versions optimized in terms of execution time building on the inherent parallelizability of the Gibbs sampler.

As anticipated, the total number of particles in the Monte Carlo data set used in this study is in line with typical charged particle multiplicities at the LHC corresponding to operating conditions as of July 2011. For this reason, the results presented in this article are a promising starting point for futher development, with a view to building dedicated software tools based on the Gibbs sampler for intensive offline analysis of individual interesting events at the LHC.

5 Acknowledgments

The author wishes to thank the High Energy Physics Group at the Department of Physics and Astronomy, University College London, for their kind hospitality that created the conditions for this study to be completed. Particular gratitude goes to Prof. Jonathan M. Butterworth for his precious comments on this project. The author also wishes to thank Prof. Trevor Sweeting at the Department of Statistical Science, University College London, for his feedback, and Dr. Alexandros Beskos at the same department for fruitful discussions. Particular gratitude also goes to the Department of Astronomy and Theoretical Physics at Lund University for their kind hospitality when this project was started, and particularly to Prof. Carsten Peterson and to Prof. Leif Lönnblad for their advice and for many fruitful discussions.

A Appendix: Toy Monte Carlo studies

Results from the Monte Carlo study described in section 3 were crosschecked on toy Monte Carlo data sets. Samples of $\sim$ 600 signal and background particles were generated according to η and p_T distributions similar to those obtained from Pythia. Particle η and p_T were generated independently: Gaussian PDFs centered at zero with standard deviations comparable to those observed in the Monte Carlo were used for η, and p_T values were generated based on polynomial PDFs in the range 2 GeV/c $< p_T <$ 5 GeV/c parametrizing the

corresponding Monte Carlo distributions.

Additional crosschecks were performed by varying both toy Monte Carlo generation parameters and seeds, in order to make sure that results did not depend on a specific parameter choice. The algorithm was also run on different numbers of particles in the input data set, with no significant changes to the results.

As mentioned when discussing the bias on mixture weights in section 3.2.1, toy Monte Carlo data sets made it possible to rule out an intrinsic bias in the supervised Gibbs sampler as far as mixture weights are concerned. The following remarks can be made with reference to the results of the supervised Gibbs sampler on the toy Monte Carlo samples:

- Particle η and p_T are statistically independent by construction, so that the correlation between the two variables in the data set used for the Monte Carlo study, which is anyway expected to have a small effect, cannot contribute to a bias on mixture weights on the toy Monte Carlo samples.

- The above-mentioned "sampling bias" due to subpopulation PDFs from high-statistics samples being "broader" than the corresponding lower-statistics distributions is not relevant when the PDFs used with the supervised Gibbs sampler come from truth information.

The above-mentioned figure 6 shows typical estimates of mixture weights obtained running the supervised Gibbs sampler on a toy Monte Carlo data set using PDFs from truth information. Results from the last 100 iterations are shown. Solid green and dashed red curves correspond to estimated fractions of signal and background particles, respectively. The solid green (dashed red) horizontal line indicates the true value for signal (background) from the toy Monte Carlo, and the dash-dot line corresponds to the initial conditions for the mixture weights. The agreement with the true weights is significantly better than the corresponding results obtained running the Gibbs sampler on the Monte Carlo data set.

References

[1] Marin, J. M., "Bayesian Modelling and Inference on Mixtures of Distributions", in: Dey, D. and Rao, C. R. (Eds.) Handbook of Statistics, Elsevier, ISBN 9780444515391, 2005.

[2] Geman, S. and Geman, D., *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** (1984) 721.

[3] Groeneboom, N. E., arXiv:0905.3823v1 [astro-ph.CO] (2009).

[4] D'Agostini, G., "Bayesian reasoning in high energy physics - Principles and applications", CERN Yellow Report 99-03, ISSN 0007-8328, ISBN 92-9083-145-6, 1999.

[5] D'Agostini, G., "Bayesian Reasoning in Data Analysis: A Critical Introduction", World Scientific Publishing Company, ISBN-10: 9812383565, ISBN-13: 978-9812383563 edition, 2003.

[6] Ciuchini, M. *et al.*, JHEP07 **013** (2001).

[7] Buckley, A. *et al.*, arXiv:hep-ph/0605048v1 (2006).

[8] Sjöstrand, T., Mrenna, S., and Skands, P., *JHEP05* **05** (2006).

[9] Sjöstrand, T., Mrenna, S., and Skands, P., *Comput. Phys. Comm.* **178** (2008).

[10] Walsh, B., "Markov Chain Monte Carlo and Gibbs Sampling", Lecture Notes for EEB 581, version 26 April 2004 (C) B. Walsh 2004, http://web.mit.edu/∼wingated/www/introductions/mcmc-gibbs-intro.pdf.

[11] Draper, D., http://users.soe.ucsc.edu/∼draper/San-Francisco-2010-notes-part1.pdf (2010).

[12] Hoecker, A. and Kartvelishvili, V., Nucl. Inst. & Meth. in Phys. Res. A **372**(3) (1996) 469-81; arXiv:hep-ph/9509307.

[13] Krzanowski, W. J. and Hand, D. J., "ROC Curves for Continuous Data", Chapman & Hall/CRC Monographs on Statistics & Applied Probability, ISBN: 978-1-4398-0021-8, 2009.

[14] Hoecker, A. *et al.*, PoS A CAT **040** (2007).

[15] Lockman, W., SLUO/LHC Workshop, July 2009.

[16] Brun, R. and Rademakers, F., Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A **389** (1997) 81-86. See also http://root.cern.ch.